Analysis of Weather Events in the USA

Ricardo Alcaraz Fraga

5/8/2020

Synopsis

Severe weather events can be harmful for public health and economy, that's why getting to know the events that can harm the most can result in designing strategies to better prepare for them, or at least, recover from the impact in a better way.
This document focuses on getting to know which events are the most harmful, for health and for the economy.

Data Processing

This analysis uses the National **Weather Service Storm Data**, which is a csv file. Let's take a look at it.

```
setwd('/home/ricardo/Escritorio/Storm/Data')
data <- read.csv('repdata_data_StormData.csv.bz2')
print(head(data))
```

```
##   STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME STATE  EVTYPE
## 1       1  4/18/1950 0:00:00     0130       CST     97     MOBILE    AL TORNADO
## 2       1  4/18/1950 0:00:00     0145       CST      3    BALDWIN    AL TORNADO
## 3       1  2/20/1951 0:00:00     1600       CST     57    FAYETTE    AL TORNADO
## 4       1   6/8/1951 0:00:00     0900       CST     89    MADISON    AL TORNADO
## 5       1 11/15/1951 0:00:00     1500       CST     43    CULLMAN    AL TORNADO
## 6       1 11/15/1951 0:00:00     2000       CST     77 LAUDERDALE    AL TORNADO
##   BGN_RANGE BGN_AZI BGN_LOCATI END_DATE END_TIME COUNTY_END COUNTYENDN
## 1         0                                               0         NA
## 2         0                                               0         NA
## 3         0                                               0         NA
## 4         0                                               0         NA
## 5         0                                               0         NA
## 6         0                                               0         NA
##   END_RANGE END_AZI END_LOCATI LENGTH WIDTH F MAG FATALITIES INJURIES PROPDMG
## 1         0                      14.0   100 3   0          0       15    25.0
## 2         0                       2.0   150 2   0          0        0     2.5
## 3         0                       0.1   123 2   0          0        2    25.0
## 4         0                       0.0   100 2   0          0        2     2.5
## 5         0                       0.0   150 2   0          0        2     2.5
## 6         0                       1.5   177 2   0          0        6     2.5
##   PROPDMGEXP CROPDMG CROPDMGEXP WFO STATEOFFIC ZONENAMES LATITUDE LONGITUDE
## 1          K       0                                         3040      8812
## 2          K       0                                         3042      8755
## 3          K       0                                         3340      8742
## 4          K       0                                         3458      8626
## 5          K       0                                         3412      8642
```

```
## 6          K       0                                          3450      8748
##   LATITUDE_E LONGITUDE_ REMARKS REFNUM
## 1       3051       8806               1
## 2          0          0               2
## 3          0          0               3
## 4          0          0               4
## 5          0          0               5
## 6          0          0               6
```

As you can see, there are many variables in the data set.

- **STATE__**: Id representing each state of the USA.
- **BGN_DATE**: Date in which the weather event happened
- **BGN_TIME**: Hour in which the weather event happened
- **TIME_ZONE**: Geographic time zone for the date and time of the weather event
- **COUNTY**: County number
- **COUNTYNAME**: County name
- **STATE**: Two letter abbreviation of the state
- **EVTYPE**: Type of event
- **BGN_RANGE**: Beginning range
- **BGN_AZI**: Beginning azimuth
- **BGN_LOCATI**: Beginning location
- **END_DATE**: End date for the event
- **END_TIME**: End time for the event
- **END_RANGE**: Range for the end
- **END_AZI**: Azimuth for the end
- **END_LOCATI**: Location where the event ended
- **LENGTH**: Length of the event
- **WIDTH**: Width of the event
- **MAG**: Magnitude
- **FATALITIES**: Lives lost in the event
- **INJURIES**: Total injuries in the event
- **PROPDMG**: Property damage (dollars)
- **PROPDMG_EXP**: Units (k, m, b) -> (thousands, millions, billions)
- **CROPDMG**: Crops damage (dollars)

First, we'll address the question about **public health**, once we've answered that question we'll, hopefully, we'll be able to use a similar methodology to answer the question about **economy**.
To address the question about **public health** we'll use the next columns:

- EVTYPE
- FATALITIES
- INJURIES

To address the question about **economy** we'll use the next columns:

- EVTYPE
- CROPDMG
- CROPDMGEXP*
- PROPDMG
- PROPDMGEXP*

The columns with a * are going to function as indicators to get the columns CROPDMG and PROPDMG, I'll explain this when we get to the economy analysis.
Other than this, I'm going to use the data like this, with no normalization or so.

Results

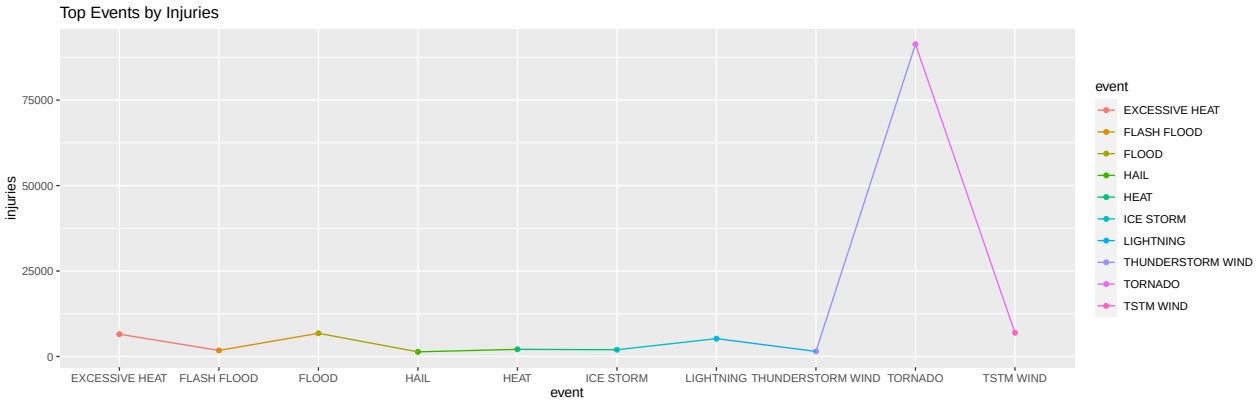
Public Health

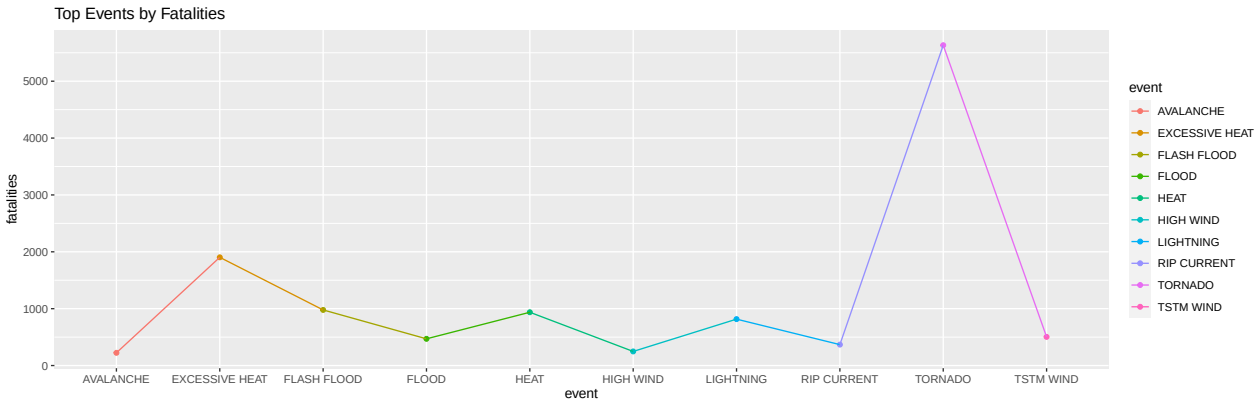
Almost immediately it comes to mind to analyze **FATALITIES** and **INJURIES** to answer this question. Let's take a look at this data grouped by the event type.

```
fatalities_injuries <- data[, c(8, 23, 24)]

event_types <- unique(fatalities_injuries$EVTYPE)

fatality <- c()
injury <- c()
for(event in event_types){
      fatality <- c(fatality, sum(fatalities_injuries[
            which(fatalities_injuries$EVTYPE == event), 'FATALITIES']))

      injury <- c(injury, sum(fatalities_injuries[
            which(fatalities_injuries$EVTYPE == event), 'INJURIES']))
}

fatalities_injuries_evtype <- data.frame(event = event_types,
                                         fatalities = fatality,
                                         injuries = injury)

print(head(fatalities_injuries_evtype), 10)
```

```
##                   event fatalities injuries
## 1               TORNADO       5633    91346
## 2             TSTM WIND        504     6957
## 3                  HAIL         15     1361
## 4         FREEZING RAIN          7       23
## 5                  SNOW          5       29
## 6 ICE STORM/FLASH FLOOD          0        2
```

There are 985 different event types and, as you can imagine, plotting all this values would be madness! So, I'm going to show the event types with the most injuries and fatalities.





It surprises no one, but tornadoes are the top in injuries and fatalities. The data set only contains this info on health, so, we can assure that **tornadoes are the events that harm the most the public health**.

Economy

Now that the question about health is answered, let's tackle the one about economy, how do we determine this? Analogously to health, we'll sum up the damage to crops and property. There is a huge problem with the data set in this matter, the columns CROPDMGEXP and PROPDMGEXP make the time for computing the total damage ridiculously high, which is a problem for our analysis, but I think I might have a solution. The _DMGEXP columns gives us the amount of zeros we have to add to the value of damage. It can have the values K, M and B. Or at least that's what the pdfs (given by them) says, but that's not true, there are more possible values, which we can see here.

```
print(unique(data$PROPDMGEXP))
```

```
##  [1] K M   B m + 0 5 6 ? 4 2 3 h 7 H - 1 8
## Levels:  - ? + 0 1 2 3 4 5 6 7 8 B h H K m M
```

```
print(unique(data$CROPDMGEXP))
```

```
## [1]   M K m B ? 0 k 2
## Levels:  ? 0 2 B k K m M
```

How do we compare damages in a reasonable computational time? We should look at values of B, as billions are way bigger than thousands and millions, so this can give us a pretty good idea about which type of events have a bigger impact on economy.

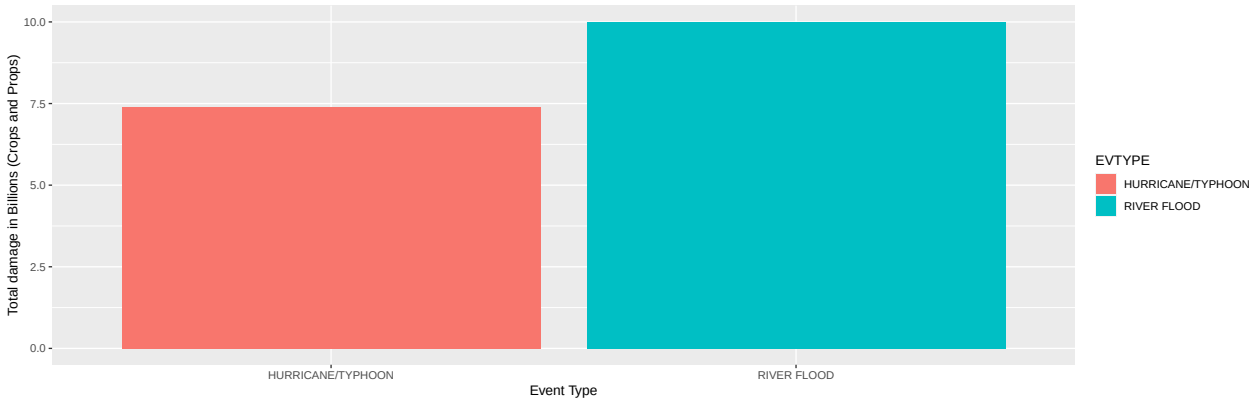
```
billions_crops <- which(data$CROPDMGEXP == 'B')
billions_props <- which(data$PROPDMGEXP == 'B')

print(length(billions_crops))
```

```
## [1] 9
```

```
print(length(billions_props))
```

```
## [1] 40
```

It seems that, speaking about damage in billions, properties have it worse than crops. Let's sum the damage for events in which both properties and crops receive damage values in billions.



From this analysis we can conclude that **RIVER FLOOD has the biggest impact on economy**.